Prospective identification of elevated circulating CDCP1 in patients years before onset of lung cancer.

Sonia Dagnino[1+], Barbara Bodinier[1+], Florence Guida[2], Karl Smith-Byrne[2], Dusan Petrovic[1,3,4], Matthew D. Whitaker[1], Therese Haugdahl Nøst[5], Claudia Agnoli[6], Domenico Palli[7], Carlotta Sacerdote[8], Salvatore Panico[9], Rosario Tumino[10], Matthias B. Schulze[11,12], Mikael Johansson[13], Pekka Keski-Rahkonen[2], Augustin Scalbert[2], Paolo Vineis[1,14], Mattias Johansson[2], Torkjel M. Sandanger[5], Roel C.H. Vermeulen[1,15], Marc Chadeau-Hyam[1,15,*]

1. MRC Centre for Environment and Health, Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, United Kingdom
2. International Agency for Research on Cancer (IARC), Lyon, France
3. Department of Epidemiology and Health Systems (DESS), University Center for General Medicine and Public Health (UNISANTE), Lausanne, Switzerland
4. Department and Division of Primary Care Medicine, University Hospital of Geneva, Geneva, Switzerland
5. Department of Community Medicine, UiT- The Arctic University of Norway, Tromsø, Norway
6. Epidemiology and Prevention Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milano.
7. Cancer Risk Factors and Life-Style Epidemiology Unit, Institute for Cancer Research, Prevention and Clinical Network - ISPRO, Florence, Italy
8. Unit of Cancer Epidemiology, Città della Salute e della Scienza University-Hospital, Turin Italy
9. Department of Clinical Medicine and Surgery, Federico II University, Naples, Italy
10. Cancer Registry and Histopathology Department, Provincial Health Authority (ASP) Ragusa, Italy
11. Department of Molecular Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke, Nuthetal, Germany
12. Institute of Nutritional Science, University of Potsdam, Nuthetal, Germany
13. Department of Radiation Sciences, Oncology, Umeå University, Umeå, Sweden
14. Italian Institute of Technology, Genova, Italy
15. Division of Environmental Epidemiology, Institute for Risk Assessment Sciences, Utrecht University, Utrecht, The Netherlands

[+] These authors contributed equally to this work

*Corresponding author:
Professor Marc Chadeau-Hyam (email: m.chadeau@imperial.ac.uk)
St Mary's Hospital, School of Public Health
Norfolk Place, W12PG London, United Kingdom

Running title: Proteins and transcripts in prospective Lung Cancer cases

Conflict of interests: The authors declare no potential conflicts of interest.

Disclaimer: Where authors are identified as personnel of the International Agency for Research on Cancer / World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy, or views of the International Agency for Research on Cancer / World Health Organization.

Abstract (229 words)

Increasing evidence points to a role for inflammation in lung carcinogenesis. A small number of circulating inflammatory proteins have been identified as showing elevated levels prior to lung cancer diagnosis, indicating the potential for prospective circulating protein concentration as a marker of early carcinogenesis. In order to identify novel markers of lung cancer risk, we measured a panel of 92 circulating inflammatory proteins in 648 pre-diagnostic blood samples from two prospective cohorts in Italy and Norway (women only). To preserve the comparability of results and protect against confounding factors, the main statistical analyses were conducted in women from both studies, with replication sought in men (Italian participants). Univariate and penalized regression models revealed for the first time higher blood levels of CDCP1 protein in cases that went on to develop lung cancer compared to controls, irrespective of time to diagnosis, smoking habits, and gender. This association was validated in an additional 450 samples. Associations were stronger for future cases of adenocarcinoma where CDCP1 showed better explanatory performance. Integrative analyses combining gene expression and protein levels CDCP1 measured in the same individuals suggested a link between CDCP1 and the expression of transcripts of LRRN3 and SEM1. Enrichment analyses indicated a potential role for CDCP1 in pathways related to cell adhesion and mobility, such as the WNT/β-catenin pathway. Overall, this study identifies lung cancer-related dysregulation of CDCP1 expression years before diagnosis.

Statement of significance:

Prospective proteomics analyses reveal an association between increased levels of circulating CDCP1 and lung carcinogenesis irrespective of smoking and years before diagnosis, and integrating gene expression indicates potential underlying mechanisms.

Keywords: Proteomics, Multi-omics, Biomarker, Inflammation, Lung Cancer, Smoking, CDCP1, EGFR

Introduction

Growing epidemiological evidence has indicated the central role of inflammation and chronic inflammation in carcinogenesis, which is now widely recognized as one of the hallmarks of cancer (1,2). Chronic inflammation is a risk factor for the induction of certain cancers, and cancer itself can induce local inflammation processes that may promote tumour proliferation and metastasis(3-5). In blood, inflammation can be measured by the abundance of cytokines and other circulating proteins, which may therefore potentially serve as biomarkers of early carcinogenesis. In particular, inflammatory markers such as the C-reactive protein (CRP) and interleukins (ILs) have previously been identified as putative cancer prognostic markers (1,6).

Lung cancer is the leading cause of cancer-related mortality worldwide . Inflammation and multiple chronic inflammatory conditions are associated with an increased risk of lung cancer. Although the potential for inflammation to exacerbate the harmful effect of smoking may partially explain this association (4), the detailed mechanisms at play remain unclear (7-9). Recent studies have identified associations linking the level of circulating inflammatory proteins and lung cancer risk (4,6,7,10,11). Increased levels of IL-6 and IL-8 have been associated with higher lung cancer risk in prospective cases less than 5 years before diagnosis (7,11) as well as after 15 years of follow-up(11). These inflammatory proteins were suspected to be involved in pathways for smoking-induced carcinogenesis (7,11,12). These studies were however based on a very limited number of assayed inflammatory markers (less than 10) (7,11) and/or on the analysis of serum from cases with a short time to diagnosis (less than 5 years)(7,12). Our study extends these analyses by including a large panel (N=92) of circulating inflammatory proteins in relation to future risk of lung cancer in participants who were healthy at baseline and were followed-up for up to 16 years. Our analyses include more than 600 prospective participants from two cohorts as a discovery set, and an additional 450 participants for validation. We also adopt an integrative approach combining

proteomic and transcriptomic data obtained from the same samples to explore pathways and molecular mechanisms related to the identified protein markers.

Materials and Methods

Study population

Plasma samples were collected in participants from two prospective cohorts within the European Prospective Investigation into Cancer and Nutrition (EPIC): EPIC-Italy (13) and the Norwegian Women and Cancer Study (NOWAC)(14,15) as already described. Details on participants are available in SI Table 1. Our study population includes 325 lung cancer cases (N=192 EPIC Italy, N=133 NOWAC) and 325 healthy controls matched on age, gender, year of recruitment, season of blood collection and centre. Due to issues with the seal of the straws in which the samples were aliquoted, two EPIC-Italy cases were excluded, leaving 323 lung cancer cases. All study participants gave written informed consent for the study. For EPIC Italy, the research was approved by the Ethics Committees at the Italian Institute of Genomic Medicine (IIGM, Turin, Italy). For NOWAC, the study was approved by the Regional Committee for Medical and Health Research Ethics in North Norway. Validation of specific result was sought in 450 additional samples including 316 (161 Cases and 155 Controls) from the EPIC cohort (Centres of Netherlands, UK, Germany and Spain)(16) and 134 from the Northern Sweden Health and Disease Study (NSHDS)(17). The characteristics of the validation dataset are reported in SI Table 2. Circulating levels of proteins from these samples were measured using the same platform as in our dataset.

Inflammatory proteins measurements

The levels of 92 inflammatory proteins were measured in citrate plasma samples by multiplex proximity extension assay with the manufacturer kit (Proseek Multiplex Inflammation I panel, Olink Bioscience, Uppsala, Sweden) using a Fluidigm Biomark reader (Fluidigm Corporation, USA), as already described previously (18). The case-control paired samples were randomized over eight 96-well plates. To evaluate the repeatability of the measurements, we included two replicates for 56

EPIC-Italy participants (N=28 cases and 28 controls) and 16 control samples with the same composition. Protein levels were expressed as normalized protein expression (Ct values with corrections for assay variation), and $\log_2$-transformed prior to statistical analysis. After verifying that the proportions of samples below the limit of detection (LoD) were similar in cases and controls, we excluded proteins with levels below the LoD in more than 30% of the samples (N=21). As replicated measurements were highly consistent (Lin's concordance correlation above 0.95), we used the average levels for statistical analyses. Values below the LoD were imputed using the QRILC algorithm for left-censored data, as implemented in the R package imputeLCMD(19,20). Finally, we removed samples that (i) did not pass the Quality Control (QC) provided by the manufacturer (N=15), (ii) showed sample conservation issues by visual inspection (N=15 EPIC-Italy samples), and (iii) were detected as outliers using the Filzmoser, Maronna and Werner algorithm for multivariate outlier detection (21)(N=40 additional samples).

Transcriptomics measurements

To better characterise the functional role of the proteins, this data was integrated with measured levels of 11,610 transcripts, available for NOWAC participants (N=222). Transcriptomics data was obtained from total RNA extraction from blood cells as previously described (22,23) and log-transformed prior to statistical analyses. Briefly, RNA was extracted from buffy coats, miRNA expression profiling was performed on an Agilent Human miRNA Microarray (Release 19.0, 8 × 60K), representing 2006 human miRNAs. Pre-processing and quality assessment of the data was performed as previously described (22,23), missing values imputation was obtained with k-nearest neighbour method.

Statistical analyses

Women are over-represented in our study population notably because the NOWAC study only includes women. We restricted our primary analyses to women participants to (i) maximise sample size while protecting against un-modelled potential gender and centre related confounding, and (ii) to ensure we could integrate gene expression data that was only available in NOWAC participants. The findings were however sought for replication in men (from the Italian cohort). For completeness, however, and when applicable, we also included as a sensitivity analysis results from the analyses performed on the full study population (i.e. pooling men and women).

Univariate models

To account for technical variability, the data was de-noised by extracting the residuals from linear mixed models where the proteins levels were modelled as the outcome and plate number and centre of recruitment were included as random intercepts in the model (24). The association between the measured levels of each of the proteins from the inflammatory panel and prospective lung cancer status was evaluated using a series of logistic regression models applied on the de-noised data. The disease status (outcome) was regressed against the protein levels, age, gender (for models on the full population only) and Body Mass Index (BMI). We expressed effect size estimates as odds ratios measuring the risk change for an increase of one standard deviation in protein levels. The strength of the association was evaluated using a likelihood ratio test comparing the fit (as measured by the likelihood) of models with that of models without the protein levels in the predictor set. Results were corrected for multiple testing using the Benjamini-Hochberg procedure controlling the False Discovery Rate below 0.05. To investigate the potential confounding role of smoking, analyses further adjusted for pack years were also conducted. We sought for validation of the main findings from our univariate analyses in an independent dataset (N=450). We used the same logistic model, which was subsequently adjusted for smoking status, the only smoking

exposure variable available for all (N=450) participants, and packyears for the (N=316) participants for whom this information was available.

Penalised regression

The inflammatory protein levels were jointly regressed against the future disease risk using logistic-LASSO (Least Absolute Shrinkage and Selection Operator) models adjusted on age and BMI (un-penalised effects)(25). To investigate the stability of our results, we computed the selection proportions of the proteins by fitting the model on 1,000 subsamples of 80% of the data(26). At each subsampling iteration, the logistic LASSO models were calibrated using 10-fold cross-validation minimising the binomial deviance using the R package glmnet (27), and subsequently used to estimate variable selection proportion across the 1,000 calibrated models. The effect of adjustment on smoking was investigated by including pack years in the set of predictors without penalising this variable. As we observed strong and inseparable cohort and gender effects, analyses were performed on women participants and validated separately on men.

Sensitivity analyses

To investigate potential subtype-specific effects, univariate and multivariate models were applied to cases from each histological subtype separately (adenocarcinoma N=91, large-cell carcinoma N=17, squamous-cell carcinoma N=26, and small-cell carcinoma N=32). As we observed strong cohort effects in our data we performed analyses in EPIC-Italy and NOWAC women separately. To account for the confounding role of smoking, the analyses were also performed in never and current smokers separately. As effect size may vary during the natural history of disease progression, we also run our models separately in cases diagnosed before and after the median time to diagnosis. Time to diagnosis is defined by the time elapsed from blood sample collection date (at recruitment) to the date at which lung cancer cases were diagnosed. The median time to diagnosis in women was 4.9

years and ranged from 1 to 16 years (inter-quartile range of 4.8 years). In these analyses, circulating levels of all assayed proteins in cases from the long and short time to diagnosis sub-groups separately were compared to those observed in the full set of controls.

ROC analyses

To evaluate the amount of disease-relevant information brought about by the proteins, we performed a series of logistic models with proteins levels as predictors and future disease risk as the outcome. Models were fitted on a training set of 80% of the total population size and performances were computed on a test set including the remaining 20% of the observation. Subsamples were controlled such that each training and test sets included the same proportion of cases, that was representative of that in the full population. The procedure was repeated 1,000 times. The results were visualised as Receiver Operating Characteristic (ROC) curves, showing the pointwise average, and a confidence region delimited by the 5^{th} and 95^{th} percentiles of the True and False Positive Rates (28).

OMICs integration

To gain insight into the functional role of the strongest association we identified, CDCP1 levels were regressed against transcript levels in linear mixed models adjusted for plate using random effects. Additionally, the associations between CDCP1 and biological pathways were investigated using individual-level gene expression data (N=11,610 transcripts matched to 11,485 unique gene symbols). Functionally relevant groups of transcripts were first identified using biological information from two large knowledgebases in Panther(29): Biological Processes and Reactome. For each database, the identified pathways (involving up to 684 genes for Biological Processes and 1,499 for Reactome) were then summarised using PCA. All Principal Components (PC) explaining more

than 5% of the group's variance were kept for analyses and used as a proxy for the biological pathway. In a second step, the PC of all biological pathways were regressed individually against CDCP1 (as the outcome) using linear models. To account for the overlap in transcript members between different pathways, the effective number of tests (ENT) was computed using a PCA on the entire set of summarised pathways and estimated as the number of PC needed to explain 90% of the variance. Results were corrected for multiple testing using the threshold in p-value p=0.05/ENT. All statistical analyses were performed in R, version 4.0.2.

Results

Descriptive analyses

The main features of the study population are summarised in Supplementary Table 1 and show that participants from both the Italian (EPIC Italy) and Norwegian (NOWAC) cohorts share similar characteristics, except smoking habits, and distribution of lung cancer histological subtypes. Particularly, we observed a slight excess of adenocarcinoma (47.4%) in NOWAC compared to EPIC Italy (44 and 39.6% for women and men respectively). Small cell carcinoma was the second most prevalent cancer in NOWAC (14.2%) whereas large cell carcinoma was more frequent in EPIC Italy (15.5% in women and 21.7% in men). To maximise the comparability of the population from NOWAC (women only) and EPIC-Italy, we restricted our main analyses to women, and, as sensitivity analysis, investigated separately data from Italian men. Results obtained from these sub-group analyses are compared to models applied on the full population and further adjusted on gender.

Of the 92 inflammatory proteins assayed in our samples, 21 were excluded due to levels falling below the limit of detection (LoD) in more than 30% of the samples, leaving 71 proteins for further analyses. The proportion of measurements below the LoD did not depend on the case control status (SI Figure 1A). For the 56 technical replicates included in our assays, Lin's concordance correlations of the measured protein levels were all above 0.95, indicating a good repeatability of the measurements (SI Figure 1B). To avoid generating results driven by outlying observations, we used

an automatic outlier detection algorithm applied to the first 5 principal components (see the methods for details) and excluded (N=59) participants from our analyses. We also excluded (N=15) samples which did not pass the quality control provided by the analysing laboratory, and (N=21) that had a default in sample vials prior to the analysis (SI Figure 2A). To correct for the nuisance variation and to reduce the potential for technical bias, the data was subsequently de-noised by extracting the residuals from linear mixed models with centre and plate ID as random intercepts (SI Figure 2B and C).

Univariate analyses reveal higher levels of protein CDCP1 in future cases

Univariate logistic regression models indicated that the circulating levels of twelve proteins: CDCP1 (OR=1.94, $p=5.49x10^{-9}$), HGF (OR=1.43, $p=6.82x10^{-4}$), IL6 (OR=1.46, $p=7.63x10^{-4}$), OSM (OR=1.41, $p=1.09x10^{-3}$), MCP1 (OR=1.38, $p=2.12x10^{-3}$), IL8 (OR=1.29, $p=3.84x10^{-3}$), VEGFA (OR=1.33, $p=5.39x10^{-3}$), CD6 (OR=1.32, $p=7.08x10^{-3}$) and CD5 (OR=1.32, $p=7.41x10^{-3}$) were associated with an increased risk of lung cancer in women after adjustment for multiple testing using the Benjamini-Hochberg procedure (FDR) (Table 1). Levels of SCF (OR=0.62, $p=1.02x10^{-5}$), TWEAK (OR=0.76, $p=6.47x10^{-3}$) and IL12B (OR=0.75, $p=6.65x10^{-3}$) were inversely associated with lung cancer risk. All these twelve proteins were also associated with exposure to tobacco smoke as measured by pack years or smoking status (SI Table 3). SCF, TWEAK and IL12B were the only proteins showing decreased levels in relation to smoking.

Associations between all proteins and future lung cancer risk were attenuated when adjusting for smoking (as measured by pack years, Table 1), and only CDCP1 (OR=1.58, $p=3.09x10^{-4}$) remained clearly associated with risk, albeit with a partly attenuated association with the risk of lung cancer independently of smoking. Analyses restricted to (N=132) women who never smoked showed that CDCP1 (OR=1.46, $p=7.78x10^{-2}$), and IL8 (OR=1.61, $p=1.69x10^{-2}$) were the most dysregulated proteins

in relation to future lung cancer, but neither survived correction for multiple testing (SI Table 4). All of the three proteins that were associated with lung cancer in the analyses restricted to current smokers survived adjustment on pack years (CDCP1, OR=2.16 $p=2.00x10^{-4}$; SCF, OR=0.53 $p=1.91x10^{-3}$; and IL6, OR=2.14 $p=6.07x10^{-4}$) (SI Table 4).

Levels of CDCP1 in men (88 cases and 88 controls) were also associated with future risk of lung cancer (OR=1.68, $p=1.51x10^{-3}$,and OR=1.86, $p=7.42x10^{-4}$ for the unadjusted and the model adjusted for smoking, respectively; SI Table 5).

Models applied to the full population and adjusted on gender yielded consistent results (SI Table 6). Eleven of the twelve proteins (all except TWEAK) identified in women were also associated with future risk of lung cancer in the full population (SI Table 6). CDCP1 was the only protein associated with the risk of lung cancer in the model adjusted for pack years in the full population and in current smokers (OR>1.83, $p<5.91x10^{-6}$) (SI Tables 6 and 7).

Validation of the association involving blood levels of CDCP1 was sought for in samples of (N=450) participants from the EPIC study (Centres of Netherlands, UK, Germany and Spain) and the NSHDS (Northern Sweden Health and Disease Study) including 225 cases and 225 healthy controls. Results consistently showed elevated levels of CDCP1 in prospective cases (SI Table 8). The associations were attenuated upon adjustment for pack years (available only for EPIC samples), and to a lesser extent for smoking status (available for both studies) but remained associated with lung cancer outcome at a nominal significance level of 0.05. We obtained consistent results in the full population, and in women and men separately.

Analyses of CDCP1 and Lung Cancer by time-to-diagnosis sub-groups and by cohort

We compared the levels of all assayed proteins in two sub-groups of cases based of the time between blood draw and clinical onset (SI Table 9) to those of all controls. We found that levels of

CDCP1 were higher in cases (and in both sub-groups of cases) than those observed in controls (SI Figure 3). Levels of CDCP1 were associated with future risk of lung cancer in both cases diagnosed before and after the median time to diagnosis (4.9 years). The association survived adjustment for smoking in the longer time to diagnosis group (OR=1.91, $p=1.67x10^{-5}$) and was borderline significant in the shorter time to diagnosis group (OR=1.35, $p=5.45x10^{-2}$).

Circulating levels of SCF were also found associated with the future risk of lung cancer irrespective of the time to diagnosis in the base model (OR<0.64, $p<3.75x10^{-4}$). Two other proteins were found associated with lung cancer risk in the shorter (OSM, OR=1.53, $p=9.08x10^{-4}$) and in the longer (MCP1, OR=1.49, $p=2.03x10^{-3}$) time to diagnosis groups, but none of these association survived correction for multiple testing in models adjusted for smoking.

In analyses by cohort, only CDCP1 was associated with lung cancer risk in both the NOWAC and EPIC Italy (OR>1.73, $p<8.04x10^{-5}$) (SI Table 10). This association was attenuated when adjusting for smoking ($p=2.26x10^{-2}$ and $2.22x10^{-3}$ in NOWAC and EPIC-Italy, respectively), and while it did not survive correction for multiple testing, it was suggestive of a consistent increased risk of lung cancer for higher levels of CDCP1 (at a nominal significance level of 0.05).

Multivariate analyses

In order to account for the complex correlation patterns across proteins in cases and controls (SI Figures 4 A and B), and in order to identify a sparse set of proteins jointly and complementarily contributing to lung cancer risk, we adopted a penalised logistic regression model using LASSO penalty to allow for variable selection. Penalised regression was coupled with stability assessment based on features selection proportion which was calculated over 1,000 sub-samples of the full population. Our logistic LASSO models consistently selected CDCP1 as well as three other proteins (selected in over 90% of sub-samples, MCP1, SCF, IL10) that were, at least partly, independently associated with lung cancer risk (Figure 1A). Our logistic LASSO also selected ST1A1, CXCL11, CD8A (with selection proportion greater than 75%), while these proteins were not found associated with

lung cancer risk in univariate analyses ($p>7.30x10^{-2}$). Possibly due to their correlation with MCP1 ($\rho>0.27$), HGF, VEGFA and CD6 were found associated lung cancer risk in the univariate models ($p<7.08x10^{-3}$) but were not frequently selected in our LASSO analyses (selection proportions ranging from 31% to 42%). Selection proportion of all proteins were attenuated in models adjusted for pack years, and only CDCP1 and IL10 remained selected with selection proportion over 80% in the adjusted model (Figure 1A).

Results of the LASSO in (N=173) men from the EPIC Italy study (SI Figure 5) highlight CDCP1 as the most frequently selected protein in both the unadjusted and smoking-adjusted models (with selection proportions of 81% and 70%, respectively). Five other proteins (SCF, CD5, CXCL10, FGF21, and AXIN1) were selected in over 60% of the sub-samples in the base model, but their selection proportion dropped below 40% in the model adjusted for smoking. These results were consistent with those obtained from the full population, where CDCP1 is the only protein with a selection proportion above 0.8 for analyses all lung cancer, adenocarcinoma, and small cell carcinoma cases (SI Figure 6 A, B and C, respectively).

Inflammatory proteins and cancer subtypes

To investigate the role of CDCP1 in association with specific subtypes and in order to account for histological heterogeneity of lung cancer, we ran our analyses on the four most common subtypes represented in our study: adenocarcinoma (N=91 cases), small-cell carcinoma (N=32), squamous cell carcinomas (N=26) and, large cell carcinoma (N=17), and compared them to all (N=201) controls. CDCP1 was found associated with the risk of adenocarcinoma (OR=1.84, $p=5.24x10^{-5}$) and small-cell carcinoma (OR=2.73, $p=7.82x10^{-5}$) in the model adjusted for pack years (Table 1). Results obtained from models applied to the full population also suggest an association between CDCP1 and adenocarcinoma (OR=1.98, $p=2.52x10^{-8}$) and small cell carcinoma (OR=4.1, $p=8.03x10^{-12}$) in both the unadjusted model and the model adjusted for pack years (SI Table 6).

Despite the small number of observations in the validation set, results also suggested an association for higher levels of CDCP1 and the risk of adenocarcinoma ($2.24x10^{-4}$, $2.28x10^{-4}$ and $1.42x10^{-2}$ for the base model and for the model adjusted for smoking status, or pack years respectively). Results were consistent but weaker in the analyses by gender (SI Table 8).

In our population, the levels of eight other inflammatory proteins were increased in future small-cell carcinoma cases (HGF, MCP1, CD6, IL18, CCL11, IL10RB, TRAIL, and CCL3, $p<0.005$). Blood levels of SCF were inversely associated with the risk of squamous-cell carcinoma (OR=0.47, $p=6.39x10^{-5}$) (Table 1). Of these, five (IL18, CCL11, IL10RB, TRAIL, and CCL3) were not associated with adenocarcinoma, large-cell or squamous-cell carcinoma ($p>0.09$).

Sub-type specific penalized regression models consistently selected CDCP1 (selection proportion of 0.99 and 0.98 in the models unadjusted and adjusted for pack years, respectively), and IL10 (selection proportion of 0.8 in the unadjusted model and 0.95 in the model adjusted on pack years, respectively) as jointly explaining the risk of adenocarcinoma (Figure 1B). For the risk of small cell carcinoma (Figure 1C), penalized regression model selected CDCP1, IL12B, and CCL11 (selection proportion>0.8) as jointly contributing to risk in the unadjusted model. CDCP1 remained highly selected (selection proportion of 0.81) in the model adjusted for smoking, while selection proportions of IL12B and CCL11 dropped below 0.1.

Quantification of the explanatory abilities of CDCP1

We conducted Receiver Operating Characteristic (ROC) analyses to quantify the abilities of the circulating proteins to discriminate between future lung cancer cases and controls (Figure 2A) in all women. The model including CDCP1 alone yielded a mean AUC of 0.65. The amount of disease-related information added by CDCP1 over and above that of pack years was modest (mean AUC of 0.74 in the model with pack years alone, and 0.75 with pack years and CDCP1). The inclusion of additional proteins selected in the LASSO (N=10 proteins with selection proportions$\geq$0.8) improved

the explanatory performance over that of the model only including CDCP1 on top of pack years (mean AUC=0.78). This suggests that selected proteins capture complementary disease-relevant information and slightly improve the discriminatory performance of pack years only.

For adenocarcinoma, CDCP1 yielded a slightly higher explanatory performance (mean AUC=0.68, Figure 2B), which was comparable to that of the model with pack years alone (mean AUC=0.69). Including both CDCP1 and pack years improved the performance of the model (mean AUC=0.73), suggesting that CDCP1 provided additional risk-relevant information to packyears for adenocarcinoma. Conversely, the risk of small cell carcinoma is more accurately explained by pack years alone (AUC=0.88), and neither CDCP1, nor the set of proteins selected by the LASSO (N=3) improved risk explanation (Figure 2C).

Correlation between CDCP1 levels and full resolution gene expression data suggest a role of the WNT signalling pathway

To better characterise the functional role of CDCP1, we explored the correlations between blood levels of CDPC1 and the levels of 11,610 transcripts previously assayed in the same NOWAC participants (N=222). Univariate linear models regressing CDCP1 levels (as the outcome) against transcript levels, identified significant associations linking levels of CDPC1 and the expression of *LRRN3* and *SEM1* (SI Figure 7) after FDR control using the Benjamini-Hochberg approach. To ease results interpretation, we defined functional groups of transcripts using the Reactome and Biological Processes (Gene Ontology) knowledgebases(30,31). We identified 1,545 and 3,600 functional groups for Reactome and Biological Processes databases respectively, and each were summarised using PCA (we included all principal components explaining at least 5% of the variance of each pathway). In models regressing these summary variables against CDCP1, three Reactome pathways were significantly associated with CDCP1 after correction for multiple testing: Initial triggering of

complement, Defective GALNT3 causes familial hyperphosphatemic tumoral calcinosis, and Deactivation of the beta-catenin transactivating complex ($p<4.17x10^{-4}$) (Figure 3A, SI Table 11A). These involved 6 to 30 transcripts, none of which were detected in univariate regressions. Using biological processes for the grouping, we identified eleven significant pathways, including: Protein localization to nucleus, Regulation of cell-cell adhesion and Regulation of chemotaxis (Figure 3B, SI Table 11B).

Discussion

We assessed the association of a panel of circulating inflammation proteins with risk of subsequent lung cancer in two prospective cohorts as training set and validated our main finding in independent samples. We adopted complementary statistical approaches, which consistently identified CDCP1 as being directly associated with future lung cancer risk irrespective of time to diagnosis and smoking habits. In our univariate models, 12 proteins from our panel were found associated with lung cancer status, including CDCP1, SCF, IL6 and IL8. Consistently with previous studies we observed increased levels of IL6 and IL8 associated with future lung cancer cases, although these association were weaker than previously described, and were attenuated when accounting for tobacco exposure (7,32). From our analyses, CDCP1 stood out due to its strong and consistent association with prospective lung cancer risk, irrespective of smoking habits (as measured by pack years) and time to diagnosis. Overall, our results point towards elevated blood levels of CDCP1 in prospective lung cancer cases compared to controls (10.6%, 8.4%, and 9.9% increase in women, men and the full population, respectively), irrespective of smoking habits and time to diagnosis.

Subtype analyses identified several differentially expressed proteins (in particular for small cell carcinoma cases). While this may indicate subtype-specific dysregulation of inflammation, these results should be taken very carefully as these are based on a very limited number of observations.

CUB domain containing protein 1, or CDCP1 is a transmembrane noncatalytic receptor involved in the loss of anchorage in epithelial cells during mitosis (33). CDCP1 has been shown to be highly expressed in different types of cancer cells and particularly human colorectal and lung cancers (34). In lung cancer, it was shown to be associated with higher proliferation, poor prognosis, survival rate and metastasis (35-38). Our findings demonstrated higher levels of circulating CDCP1 many years prior to lung cancer diagnosis, suggesting that CDCP1 is indicative of mechanisms important for lung cancer aetiology, in addition to its potential role as prognostic marker. Our ROC analyses indicated modest explanatory abilities of CDCP1 for lung cancer and its main histological subtypes. When combined with smoking, CDCP1 yielded (moderate) improvements in the explanation of lung cancer, suggesting that some of the CDCP1-lung cancer association we observe is not directly related to smoking and explains some other aspect of the lung cancer risk. Blood levels of CDCP1 may therefore have the potential to inform on the mechanisms of smoking-related and smoking unrelated lung carcinogenesis.

Pathways related to CDPC1 and pre-diagnostic lung cancer have been poorly described to date. To better understand the potential pathways linking CDPC1 and early carcinogenesis, we integrated gene expression data measured in the same individuals. We detected two transcripts associated with CDPC1: LRRN3 and (Leucine-rich repeat neuronal protein 3) and SEM1 (26S Proteasome Complex Subunit). In previous work, we identified LRRN3 to be positively associated with smoking exposure (39). Its positive association with CDPC1 suggest a potential implication of CDCP1 in smoking-induced lung cancer related pathways. SEM1 has, to our knowledge, not been reported as linked to lung cancer or expression of CDCP1.

Reactome enrichment analysis revealed an association between CDCP1 and deactivation of the β-catenin transactivating complex. WNT/β-catenin's inappropriate activation has been linked to a wide range of cancers (40,41). Beta-catenin forms a complex with TCF transcription factor family resulting in the activation of genes implicated in tumour development. In vitro, WNT/β-catenin signalling has

been identified as a critical pathway in human lung carcinogenesis (42). Here we show that higher levels of CDCP1 are negatively associated with the β-catenin deactivation pathway. In accordance with our findings, recent work in colorectal cancer cell lines have shown that CDCP1 is an important regulator of WNT signalling, and that similar to what we observe for lung cancer, elevated levels of CDCP1 were predictive of colorectal cancer(43). In accordance with these observations, pathway analysis with the PANTHER biological processes database also indicated that CDCP1 was associated with protein localization in the nucleus, as observed by Hu et al.(43). An additional interesting pathway associated with CDCP1 was the cell-cell adhesion. Dong et al. observed that a disruption of CDCP1 in vitro was associated with an interference in EGF/EGFR (Epidermal growth factor receptor) induced cell migration and suggested CDCP1 as a potential target for EGFR driven cancers(44). EGFR is an important therapeutic target for lung cancer, as over 60% of non-small cell lung carcinomas express EGFR. In lung cancer metastasis it was shown that EGF stimulation increases CDCP1 expression and that EGFR inhibitor reduces the level of CDCP1 in lung cancer cells(45).

Strengths of our study include the validation of our association involving CDCP1 in three separate cohorts (EPIC, NOWAC, NSHDS) and with two distinct statistical methods, which enabled us to demonstrate the robust link between pre-diagnostically measured CDCP1 and risk of subsequent lung cancer, triangulate the evidence linking prospective blood levels of CDCP1, and future risk of lung cancer. Our work is also the first study to use pre-diagnostic data on both a broad panel of inflammatory proteins and gene expression in hundreds of lung cancer cases from the same population for integrated analyses, allowing for a more comprehensive assessment of potential mechanisms underlying the risk associations.

The limitations of our study include the relatively small sample size for the validation of our results in men only, as well as the analysis of histological subtypes. In addition, our measures of inflammatory proteins were cross-sectional and other prospective studies with multiple measures on participants would be instrumental to further investigate our findings. Information on EGFR mutation status was

not available for our cohort, and future studies investigating the role of CDPC1 in lung cancer would benefit from the inclusion of this information.

The survival of patients with lung cancer is highly dependent on accurate and early diagnosis. The United States National Lung Screening trial suggested that diagnosis by low-dose computed tomography could reduce up to 20% of lung cancer mortality, but that it is also associated with a high level of false positives, resulting in a great number of potentially benign cases to unnecessary and costly follow-up(46). The identification of early biomarkers of susceptibility could significantly improve diagnosis by improving the identification of individuals at high risk who are more likely to benefit from screening(47). The development of OMICs technology has opened new grounds for biomarker identification. Quantitative proteomics provides different protein abundance for samples from control and cases, allowing the identification of biomarkers, pathways perturbations and molecular interactions (48). Our study suggests that circulating serum levels of CDCP1 provides additional information on future lung cancer risk, over and above that afforded by information on tobacco exposure and may therefore help in the identification of molecular pathways involved in lung carcinogenesis, years before diagnosis.

Availability of data and materials

The data in EPIC that support the findings of this study are available from the corresponding author upon reasonable request. The NOWAC data cannot be shared publicly because of local and national ethical and security policy. Data access for researchers will be conditional on adherence to both the data access procedures of the Norwegian women and Cancer Cohort and the UiT The Arctic University of Norway (contact via Torkjel Sandanger, torkjel.sandanger@uit.no, Tonje Braaten tonje.braaten@uit.no, and Arne Bastian Wiik, arne.b.wiik@uit.no) in addition to the local ethical committee.

Acknowledgments

This work was supported by Cancer Research UK Population Research Committee 'Mechanomics' project grant (Grant #22184 to MC-H). The NOWAC post-genome cohort study was funded by the ERC advanced grant; Transcriptomics in Cancer Epidemiology (ERC-2008-AdG-232997). MC-H, FG, KS-B, THN, MJ, and TS acknowledge support from the Research Council of Norway (Id-Lung project FRIPRO 262111 to TS). This research was supported by Institut National Du Cancer (France, PI: Mattias Johansson) and Cancerforskningsfonden i Norrland (Sweden, PI: Mikael Johansson). MC-H, RV, acknowledge support from the H2020-EXPANSE project (Horizon 2020 grant No 874627 to RV). SD acknowledges support to Horizon 2020 Marie Skłodowska-Curie fellowship EXACT Identifying biomarkers of EXposure leading to Lung Cancer with AdduCTomics (Grant # 708392 to SD). BB received a PhD studentship from the MRC Centre for Environment and Health. RT acknowledges A.I.R.E. - O.N.L.U.S. Ragusa Italy. EPIC-Italy was funded by the Italian Association for Research on Cancer (AIRC). The EPIC-Norfolk study (DOI 10.22025/2019.10.105.00004) has received funding from the Medical Research Council (MR/N003284/1 and MC-UU_12015/1) and Cancer Research UK (C864/A14136). We are grateful to all the participants who have been part of the project and to the many members of the study teams at the University of Cambridge who have enabled this research. Authors would like to acknowledge principal investigators of the EPIC cohort for allowing validation

of data in their respective cohorts, namely Dr Rudolf Kaaks for EPIC Heidelberg, Dr Antonio Agudo for EPIC Spain, and Dr Nick Wareham for EPIC Norfolk.

Author contributions

SD and BB are joined first authors. MC-H, SD, BB conceived the study and drafted the manuscript. SD, BB, FG and KSB performed the statistical analyses. MC-H, supervised the analyses and together with RV drafted the analytical plan. EPIC Italy data was provided by PV, CA, DP, CS, SP, RT. NOWAC data was provided and curated by THN and TMS. EPIC and NSHDS data were provided by KSB, FG, MJ and MJ. THN, TMS, RV, PKR, AS provided insights into the study design, results interpretation and revised the manuscript. All authors revised the manuscript for important intellectual content and approved the submission of the manuscript.

Supplementary Information accompanies this paper.

References

1. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell **2011**;144:646-74
2. Hussain SP, Hofseth LJ, Harris CC. Radical causes of cancer. Nature reviews Cancer **2003**;3:276-85
3. Balkwill FR, Mantovani A. Cancer-related inflammation: common themes and therapeutic opportunities. Seminars in cancer biology **2012**;22:33-40
4. Engels EA. Inflammation in the development of lung cancer: epidemiological evidence. Expert review of anticancer therapy **2008**;8:605-15
5. Mantovani A, Allavena P, Sica A, Balkwill F. Cancer-related inflammation. Nature **2008**;454:436-44
6. Muller DC, Larose TL, Hodge A, Guida F, Langhammer A, Grankvist K*, et al.* Circulating high sensitivity C reactive protein concentrations and risk of lung cancer: nested case-control study within Lung Cancer Cohort Consortium. **2019**;364:k4981
7. Pine SR, Mechanic LE, Enewold L, Chaturvedi AK, Katki HA, Zheng YL*, et al.* Increased levels of circulating interleukin 6, interleukin 8, C-reactive protein, and risk of lung cancer. Journal of the National Cancer Institute **2011**;103:1112-22
8. Zhou B, Liu J, Wang ZM, Xi T. C-reactive protein, interleukin 6 and lung cancer risk: a meta-analysis. PLoS One **2012**;7:e43075
9. Munn LL. Cancer and inflammation. Wiley interdisciplinary reviews Systems biology and medicine **2017**;9
10. Guida F, Sun N, Bantis LE, Muller DC, Li P, Taguchi A*, et al.* Assessment of Lung Cancer Risk on the Basis of a Biomarker Panel of Circulating Proteins. JAMA oncology **2018**;4:e182078
11. Brenner DR, Fanidi A, Grankvist K, Muller DC, Brennan P, Manjer J*, et al.* Inflammatory Cytokines and Lung Cancer Risk in 3 Prospective Studies. Am J Epidemiol **2017**;185:86-95
12. Shiels MS, Katki HA, Hildesheim A, Pfeiffer RM, Engels EA, Williams M*, et al.* Circulating Inflammation Markers, Risk of Lung Cancer, and Utility for Risk Stratification. Journal of the National Cancer Institute **2015**;107
13. Palli D, Berrino F, Vineis P, Tumino R, Panico S, Masala G*, et al.* A molecular epidemiology project on diet and cancer: the EPIC-Italy Prospective Study. Design and baseline characteristics of participants. Tumori **2003**;89:586-93
14. Dumeaux V, Børresen-Dale AL, Frantzen JO, Kumle M, Kristensen VN, Lund E. Gene expression analyses in breast cancer epidemiology: the Norwegian Women and Cancer postgenome cohort study. Breast cancer research : BCR **2008**;10:R13
15. Lund E, Dumeaux V, Braaten T, Hjartåker A, Engeset D, Skeie G*, et al.* Cohort profile: The Norwegian Women and Cancer Study--NOWAC--Kvinner og kreft. Int J Epidemiol **2008**;37:36-41
16. Riboli E, Kaaks R. The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. Int J Epidemiol **1997**;26 Suppl 1:S6-14
17. Hallmans G, Agren A, Johansson G, Johansson A, Stegmayr B, Jansson JH*, et al.* Cardiovascular disease and diabetes in the Northern Sweden Health and Disease Study Cohort - evaluation of risk factors and their interactions. Scand J Public Health Suppl **2003**;61:18-24
18. Assarsson E, Lundberg M, Holmquist G, Björkesten J, Thorsen SB, Ekman D*, et al.* Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. PLoS One **2014**;9:e95192
19. Lazar C. imputeLCMD: A Collection of Methods for Left-CensoredMissing Data Imputation. . R package, version 20
20. Lazar C, Gatto L, Ferro M, Bruley C, Burger T. Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. Journal of proteome research **2016**;15:1116-25

21. Filzmoser P, Maronna R, Werner M. Outlier identification in high dimensions. Computational Statistics & Data Analysis **2008**;52:1694-711
22. Georgiadis P, Hebels DG, Valavanis I, Liampa I, Bergdahl IA, Johansson A*, et al.* Omics for prediction of environmental health effects: Blood leukocyte-based cross-omic profiling reliably predicts diseases associated with tobacco smoking. Scientific reports **2016**;6:20544
23. Hebels DG, Georgiadis P, Keun HC, Athersuch TJ, Vineis P, Vermeulen R*, et al.* Performance in omics analyses of blood samples in long-term storage: opportunities for the exploitation of existing biobanks in environmental health research. Environmental health perspectives **2013**;121:480-7
24. Chadeau-Hyam M, Campanella G, Jombart T, Bottolo L, Portengen L, Vineis P*, et al.* Deciphering the complex: methodological overview of statistical models to derive OMICS-based biomarkers. Environmental and molecular mutagenesis **2013**;54:542-57
25. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw **2010**;33:1-22
26. Vermeulen R, Saberi Hosnijeh F, Bodinier B, Portengen L, Liquet B, Garrido-Manriquez J*, et al.* Pre-diagnostic blood immune markers, incidence and progression of B-cell lymphoma and multiple myeloma: Univariate and functionally informed multivariate analyses. International Journal of Cancer **2018**;143:1335-47
27. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw **2010**;33:1-22
28. Dagnino S, Bodinier B, Grigoryan H, Rappaport SM, Karimi M, Guida F*, et al.* Agnostic Cys34-albumin adductomics and DNA methylation: implication of N-acetylcysteine in lung carcinogenesis years before diagnosis. International journal of cancer Journal international du cancer **2019**
29. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R*, et al.* PANTHER: a library of protein families and subfamilies indexed by function. Genome research **2003**;13:2129-41
30. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic acids research **2018**;47:D419-D26
31. Mi H, Thomas P. PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. Methods in molecular biology (Clifton, NJ) **2009**;563:123-40
32. Pine SR, Mechanic LE, Enewold L, Bowman ED, Ryan BM, Cote ML*, et al.* Differential Serum Cytokine Levels and Risk of Lung Cancer Between African and European Americans. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology **2016**;25:488-97
33. Hooper JD, Zijlstra A, Aimes RT, Liang H, Claassen GF, Tarin D*, et al.* Subtractive immunization using highly metastatic human tumor cells identifies SIMA135/CDCP1, a 135 kDa cell surface phosphorylated glycoprotein antigen. Oncogene **2003**;22:1783-94
34. Scherl-Mostageer M, Sommergruber W, Abseher R, Hauptmann R, Ambros P, Schweifer N. Identification of a novel gene, CDCP1, overexpressed in human colorectal cancer. Oncogene **2001**;20:4402-8
35. Ikeda J, Oda T, Inoue M, Uekita T, Sakai R, Okumura M*, et al.* Expression of CUB domain containing protein (CDCP1) is correlated with prognosis and survival of patients with adenocarcinoma of lung. Cancer science **2009**;100:429-33
36. Uekita T, Fujii S, Miyazawa Y, Iwakawa R, Narisawa-Saito M, Nakashima K*, et al.* Oncogenic Ras/ERK signaling activates CDCP1 to promote tumor invasion and metastasis. Molecular cancer research : MCR **2014**;12:1449-59
37. Uekita T, Sakai R. Roles of CUB domain-containing protein 1 signaling in cancer invasion and metastasis. Cancer science **2011**;102:1943-8

38. Zeng XJ, Wu YH, Luo M, Cong PG, Yu H. Inhibition of pulmonary carcinoma proliferation or metastasis of miR-218 via down-regulating CDCP1 expression. European review for medical and pharmacological sciences **2017**;21:1502-8
39. Guida F, Sandanger TM, Castagne R, Campanella G, Polidoro S, Palli D*, et al.* Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. Human molecular genetics **2015**;24:2349-59
40. Clevers H, Nusse R. Wnt/β-Catenin Signaling and Disease. Cell **2012**;149:1192-205
41. Polakis P. Wnt signaling in cancer. Cold Spring Harb Perspect Biol **2012**;4
42. Yuan D, Liu L, Gu D. Transcriptional regulation of livin by beta-catenin/TCF signaling in human lung cancer cell lines. Mol Cell Biochem **2007**;306:171-8
43. He Y, Davies CM, Harrington BS, Hellmers L, Sheng Y, Broomfield A*, et al.* CDCP1 enhances Wnt signaling in colorectal cancer promoting nuclear localization of β-catenin and E-cadherin. Oncogene **2020**;39:219-33
44. Dong Y, He Y, de Boer L, Stack MS, Lumley JW, Clements JA*, et al.* The cell surface glycoprotein CUB domain-containing protein 1 (CDCP1) contributes to epidermal growth factor receptor-mediated cell migration. J Biol Chem **2012**;287:9792-803
45. Chiu KL, Lin YS, Kuo TT, Lo CC, Huang YK, Chang HF*, et al.* ADAM9 enhances CDCP1 by inhibiting miR-1 through EGFR signaling activation in lung cancer metastasis. Oncotarget **2017**;8:47365-78
46. Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM*, et al.* Reduced lung-cancer mortality with low-dose computed tomographic screening. The New England journal of medicine **2011**;365:395-409
47. Kovalchik SA, Tammemagi M, Berg CD, Caporaso NE, Riley TL, Korch M*, et al.* Targeting of low-dose CT screening according to the risk of lung-cancer death. The New England journal of medicine **2013**;369:245-54
48. Cheung CHY, Juan H-F. Quantitative proteomics in lung cancer. Journal of Biomedical Science **2017**;24:37

Table 1. Logistic regression models with future disease status as the outcome and individual protein levels as predictor in Women. Models are adjusted on age and BMI (A). Centre and plate effects are removed from the data by taking the residuals from linear mixed models with protein levels as the outcome and centre and plate as random intercepts. Models further adjusted on packyears are also reported (B). The p-values of association with future disease status are derived from likelihood ratio tests comparing the fit of the model with to that of the model without protein levels in the set of predictors. Results are presented for pooled lung cancer and for each histological subtype for proteins found associated at least once with one lung cancer subtype considered after Benjamini-Hochberg correction for multiple testing.

A	All LC (N=397)		Adenocarcinoma (N=292)		Small-cell carcinoma (N=233)		Large-cell carcinoma (N=218)		Squamous-cell carcinoma (N=227)	
	OR	p-value	OR	p-value	OR	p-value	OR	p-value	OR	p-value
CDCP1	1.95	**5.49e-09**	2.41	**2.10e-08**	4.20	**1.88e-09**	1.08	8.24e-01	1.68	4.18e-02
SCF	0.63	**1.02e-05**	0.73	2.21e-02	0.58	7.22e-03	0.59	5.67e-02	0.48	**6.39e-05**
HGF	1.41	**6.82e-04**	1.39	9.05e-03	1.60	**5.34e-03**	1.06	8.32e-01	1.49	2.76e-02
IL6	1.44	**7.63e-04**	1.26	6.49e-02	1.43	1.64e-02	1.38	8.65e-02	1.40	3.77e-02
OSM	1.39	**1.09e-03**	1.39	9.00e-03	1.49	3.25e-02	0.90	6.89e-01	1.66	9.44e-03
MCP1	1.36	**2.12e-03**	1.22	1.25e-01	1.84	**2.21e-03**	1.62	5.95e-02	1.46	8.80e-02
IL8	1.30	**3.84e-03**	1.27	4.47e-02	1.31	7.85e-02	1.13	6.24e-01	1.45	2.06e-02
VEGFA	1.32	**5.39e-03**	1.16	2.42e-01	1.61	8.78e-03	0.90	6.98e-01	1.51	2.83e-02
TWEAK	0.76	**6.47e-03**	0.79	6.64e-02	0.67	3.63e-02	1.01	9.66e-01	0.56	6.60e-03
IL12B	0.76	**6.65e-03**	0.84	1.89e-01	0.62	1.45e-02	0.73	2.42e-01	0.58	1.08e-02
CD6	1.30	**7.08e-03**	1.14	2.87e-01	1.80	**1.57e-03**	1.00	9.97e-01	1.59	2.06e-02
CD5	1.31	**7.41e-03**	1.34	1.93e-02	1.58	2.37e-02	1.12	6.88e-01	1.24	3.07e-01
IL18	1.27	1.26e-02	1.17	2.06e-01	1.98	**2.82e-04**	0.85	5.41e-01	1.24	2.96e-01
CCL11	1.30	1.62e-02	1.12	4.18e-01	2.35	**1.45e-04**	1.43	2.18e-01	1.49	9.01e-02
IL10RB	1.18	9.75e-02	1.13	3.54e-01	1.99	**1.24e-03**	0.85	5.43e-01	0.81	3.24e-01
TRAIL	1.17	1.08e-01	1.22	1.18e-01	1.78	**4.15e-03**	0.88	6.09e-01	0.89	5.96e-01
CCL3	1.18	1.54e-01	0.91	5.50e-01	1.60	**6.19e-03**	1.13	6.48e-01	1.16	5.38e-01

B	All LC (N=388)		Adenocarcinoma (N=286)		Small-cell carcinoma (N=227)		Large-cell carcinoma (N=214)		Squamous-cell carcinoma (N=223)	
	OR	p-value	OR	p-value	OR	p-value	OR	p-value	OR	p-value
CDCP1	1.59	**3.09e-04**	1.95	**5.24e-05**	3.03	**7.82e-05**	1.00	9.98e-01	1.56	9.43e-02
SCF	0.78	3.94e-02	0.88	3.92e-01	0.75	2.62e-01	0.65	1.29e-01	0.56	4.15e-03
HGF	1.19	1.19e-01	1.23	1.29e-01	1.37	9.93e-02	1.01	9.80e-01	1.33	1.37e-01
IL6	1.28	3.27e-02	1.15	3.00e-01	1.42	6.20e-02	1.38	9.94e-02	1.42	4.94e-02
OSM	1.24	5.98e-02	1.26	9.07e-02	1.39	1.27e-01	0.85	5.52e-01	1.53	4.30e-02
MCP1	1.23	6.62e-02	1.12	4.19e-01	1.52	6.28e-02	1.56	8.54e-02	1.38	1.71e-01
IL8	1.29	1.16e-02	1.26	7.10e-02	1.32	1.04e-01	1.12	6.32e-01	1.27	1.56e-01
VEGFA	1.21	9.00e-02	1.09	5.26e-01	1.60	2.31e-02	0.93	7.83e-01	1.54	3.14e-02
TWEAK	0.92	4.94e-01	0.91	5.08e-01	0.99	9.79e-01	1.16	5.86e-01	0.67	1.02e-01
IL12B	0.91	4.33e-01	1.04	8.11e-01	0.83	4.33e-01	0.80	4.36e-01	0.72	1.72e-01
CD6	1.15	1.99e-01	1.02	8.76e-01	1.59	4.63e-02	0.95	8.52e-01	1.47	9.36e-02
CD5	1.16	1.72e-01	1.20	1.77e-01	1.38	2.07e-01	1.08	7.88e-01	1.20	4.32e-01
IL18	1.14	2.26e-01	1.07	6.22e-01	1.76	1.47e-02	0.80	4.22e-01	1.16	5.11e-01
CCL11	1.05	7.04e-01	0.92	6.17e-01	1.90	2.84e-02	1.37	2.95e-01	1.33	2.89e-01
IL10RB	1.17	1.72e-01	1.14	3.57e-01	2.15	4.88e-03	0.85	5.67e-01	0.97	8.92e-01
TRAIL	1.06	6.17e-01	1.11	4.34e-01	1.79	1.74e-02	0.84	5.09e-01	0.91	6.62e-01
CCL3	1.06	6.29e-01	0.86	3.94e-01	1.52	2.42e-02	1.08	7.63e-01	1.11	6.55e-01

Figure Legends:

Figure 1. Stability analyses of the logistic-LASSO models investigating the association between the 71 inflammatory proteins and future lung cancer status in women. Centre and plate effects are removed from the data by taking the residuals from preliminary linear mixed models with protein levels as the outcome and centre and plate as random intercepts. The LASSO models are adjusted on age, BMI (in blue, base model) and further adjusted on pack years (in red, model adjusted on pack years) by incorporating these covariates in the model without penalisation. Selection proportions of individual proteins are computed over 1,000 random sub-samples of 80% of the sample size and ensuring that the proportion of cases and controls is kept constant in each subsample (left panel). The penalty parameter (lambda) of the models is calibrated at each subsampling iteration using M-fold cross-validation (M=10) to minimise model deviance. Selection proportions in LASSO models are compared to the strength of association in univariate logistic models, as measured by their p-value (right panel). Analyses are conducted in participants with complete data on age, BMI and pack years. Results are presented for all lung cancer cases (A, N=191 cases) and for each subtype separately (B: Adenocarcinoma, N=89 cases, C: small-cell carcinoma, N=30 cases) compared to the 197 healthy controls.

Figure 2. Receiver Operating Characteristics (ROC) curves for logistic models for women only including (i) pack years (dark red), (ii) blood levels of CDCP1 (beige), (iii) and both CDCP1 and pack years (dark green) as predictors, as well as for a logistic-LASSO models including (iv) all inflammatory proteins (cyan) and (v) additionally pack years (dark blue). Models were fitted on a training set of 80% of the data and the performance metrics are calculated on the remaining 20% of the data (test set). The procedure was repeated 1,000 times. The logistic-LASSO model was calibrated at each iteration using 10-fold cross-validation minimising the binomial deviance. For each model, the pointwise average of the performance metrics is represented (bold line) and the area is defined by their 5^{th} and 95^{th} percentiles. Analyses are conducted in participants with complete data on age, BMI and pack years. Results are presented for pooled lung cancer cases (A, N=191 cases), adenocarcinoma (B, N=89 cases) and small-cell carcinoma (C, N=30 cases) separately. All controls (N=197) were kept for all analyses. The average and 5^{th} and 95^{th} percentiles of the Area Under the Curve (AUC) are reported in the legend.

Figure 3. Volcano plots showing the associations between CDCP1 and biological pathways, as measured by groups of transcripts defined using the Reactome (A) and Biological Processes (B) knowledgebases and summarised using the scores from Principal Components explaining more than 5% of the functional group's variance in PCA. The biological pathways significantly associated with CDCP1 after correction for multiple testing using the Effective Number of Tests (ENT=109 for the Reactome, and ENT=140 for Biological Processes) and their gene members can be visualised in the heatmaps. Other scores of significantly associated functional groups are also coloured in the Volcano plots.

Figure 1

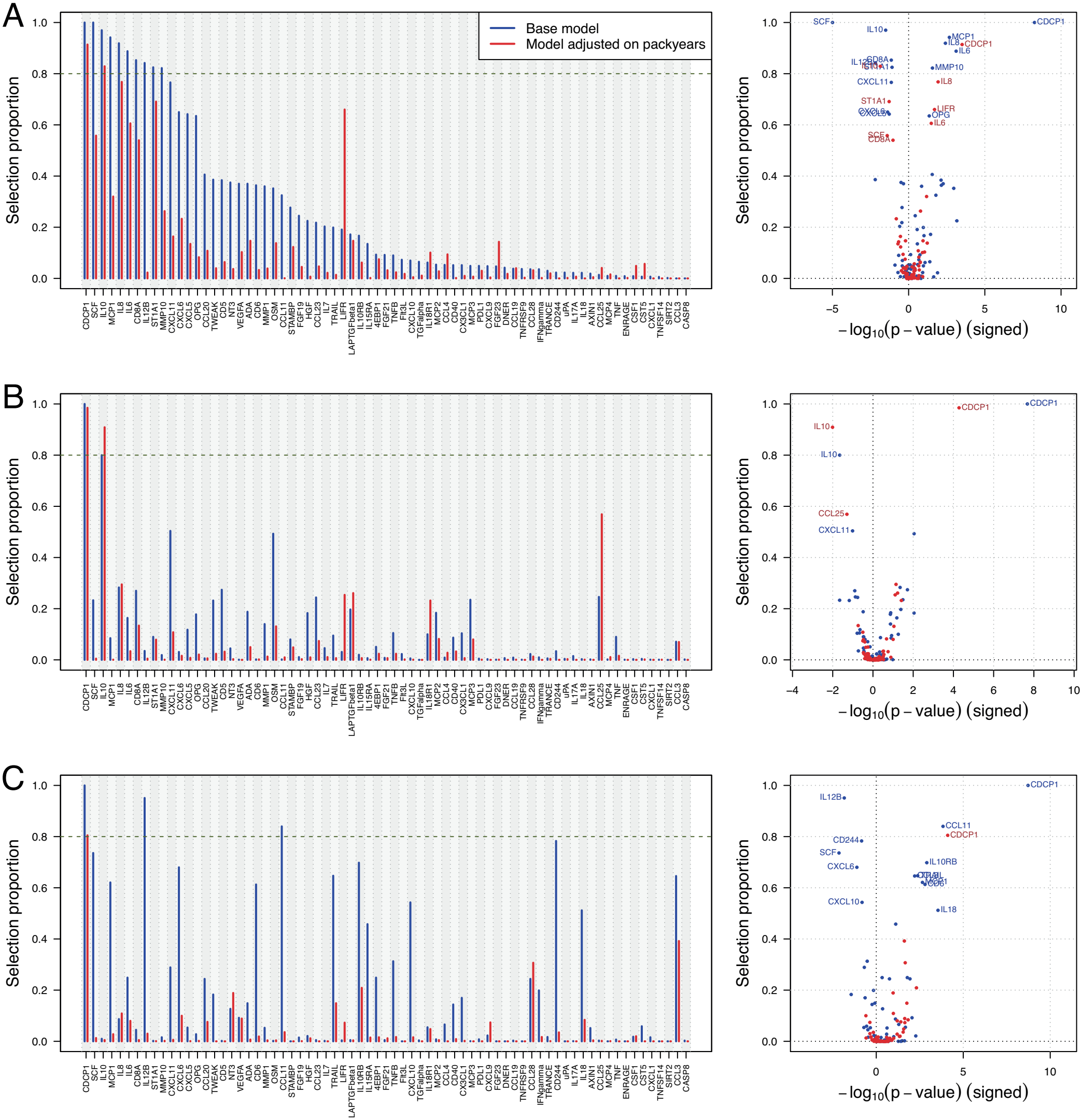

A
B
C
Base model
Model adjusted on packyears
Selection proportion
1.0
0.8
0.6
0.4
0.2
0.0
CDCP1 SCF IL10 MCP1 IL8 IL6 CD8A IL12B ST1A1 MMP10 CXCL11 CXCL6 CXCL5 OPG CCL20 TWEAK CD5 NT3 VEGFA ADA CD6 MMP1 OSM CCL11 STAMBP FGF19 HGF CCL23 IL7 TRAIL LIFR LAPTGFbeta1 IL10RB IL15RA 4EBP1 FGF21 TNFB Flt3L CXCL10 TGFalpha IL18R1 MCP2 CCL4 CD40 CX3CL1 MCP3 PDL1 CXCL9 FGF23 DNER CCL19 TNFRSF9 CCL28 IFNgamma TRANCE CD244 uPA IL17A IL18 AXIN1 CCL25 MCP4 TNF ENRAGE CSF1 CST5 CXCL1 TNFSF14 SIRT2 CCL3 CASP8
SCF
IL10
CDCP1
MCP1
IL8
IL6
CD8A
IL12B
ST1A1
CXCL11
MMP10
IL8
LIFR
OPG
IL6
SCF
CD8A
CCL25
CXCL11
IL10
IL10
CDCP1
CDCP1
IL12B
CCL11
CDCP1
CD244
SCF
CXCL6
IL10RB
CXCL10
IL18
CDCP1
−5
0
5
10
−4
−2
0
2
4
6
8
10
−log10(p − value) (signed)

Figure 2

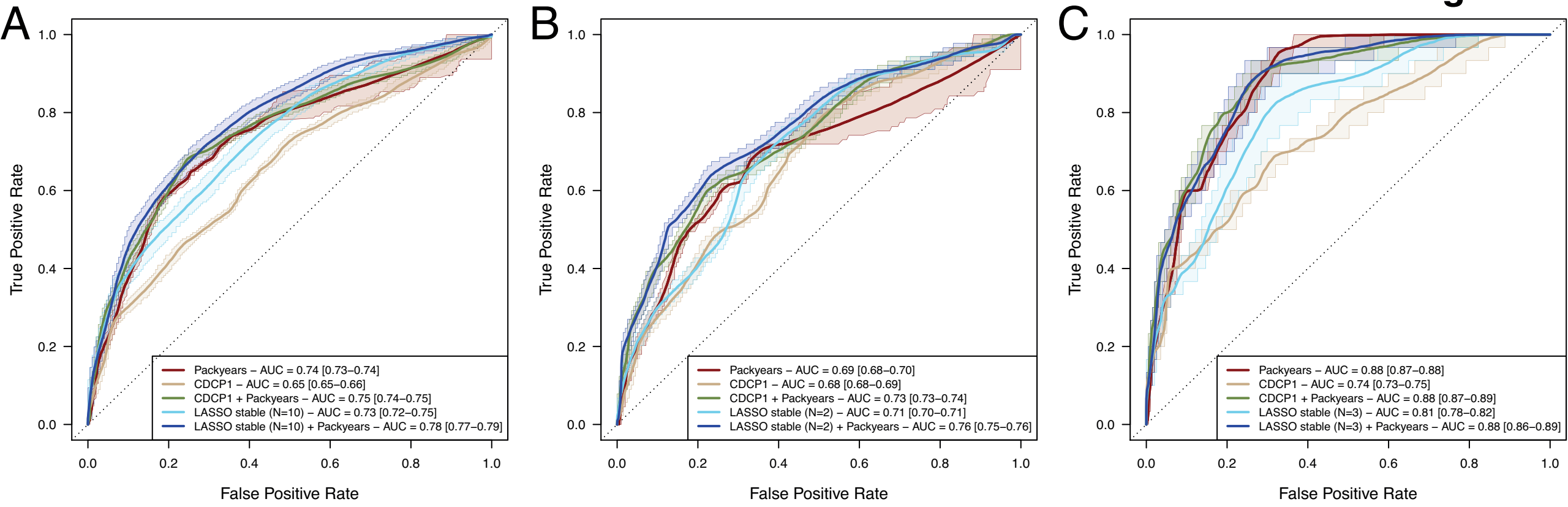
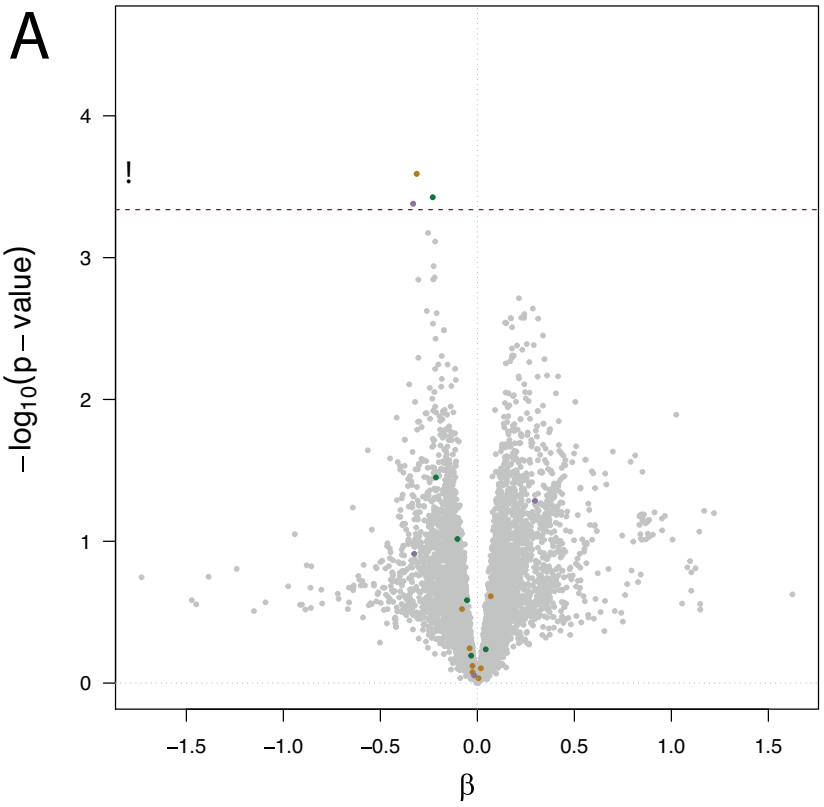
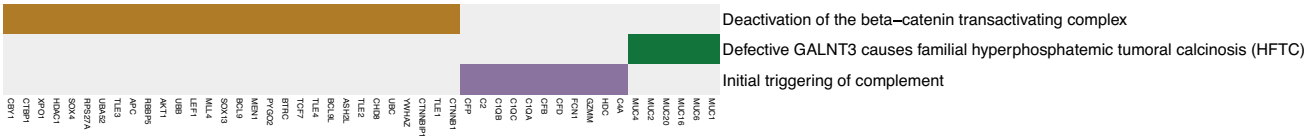

A
True Positive Rate
False Positive Rate
Packyears – AUC = 0.74 [0.73–0.74]
CDCP1 – AUC = 0.65 [0.65–0.66]
CDCP1 + Packyears – AUC = 0.75 [0.74–0.75]
LASSO stable (N=10) – AUC = 0.73 [0.72–0.75]
LASSO stable (N=10) + Packyears – AUC = 0.78 [0.77–0.79]
B
True Positive Rate
False Positive Rate
Packyears – AUC = 0.69 [0.68–0.70]
CDCP1 – AUC = 0.68 [0.68–0.69]
CDCP1 + Packyears – AUC = 0.73 [0.73–0.74]
LASSO stable (N=2) – AUC = 0.71 [0.70–0.71]
LASSO stable (N=2) + Packyears – AUC = 0.76 [0.75–0.76]
C
True Positive Rate
False Positive Rate
Packyears – AUC = 0.88 [0.87–0.88]
CDCP1 – AUC = 0.74 [0.73–0.75]
CDCP1 + Packyears – AUC = 0.88 [0.87–0.89]
LASSO stable (N=3) – AUC = 0.81 [0.78–0.82]
LASSO stable (N=3) + Packyears – AUC = 0.88 [0.86–0.89]

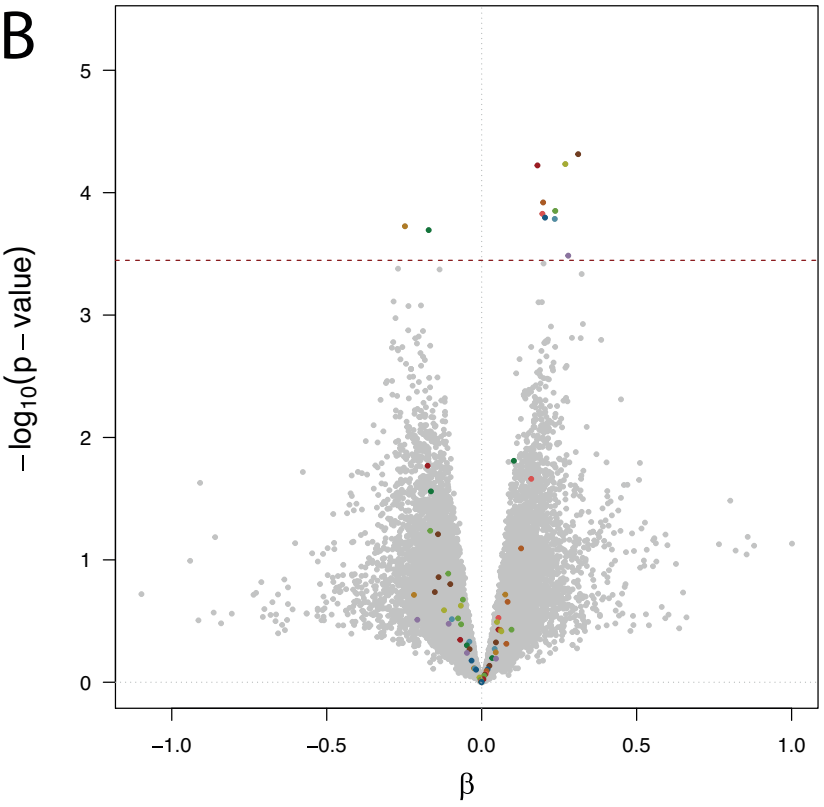
Figure 3

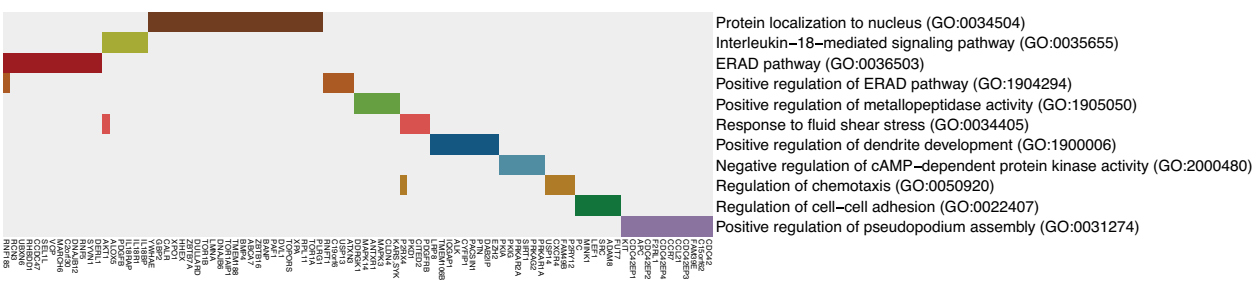
A


−log10(p−value)
β
!


Deactivation of the beta−catenin transactivating complex
Defective GALNT3 causes familial hyperphosphatemic tumoral calcinosis (HFTC)
Initial triggering of complement

B


−log10(p−value)
β


Protein localization to nucleus (GO:0034504)
Interleukin−18−mediated signaling pathway (GO:0035655)
ERAD pathway (GO:0036503)
Positive regulation of ERAD pathway (GO:1904294)
Positive regulation of metallopeptidase activity (GO:1905050)
Response to fluid shear stress (GO:0034405)
Positive regulation of dendrite development (GO:1900006)
Negative regulation of cAMP−dependent protein kinase activity (GO:2000480)
Regulation of chemotaxis (GO:0050920)
Regulation of cell−cell adhesion (GO:0022407)
Positive regulation of pseudopodium assembly (GO:0031274)